| REPORT DOCUMENTATION PAGE | *Form Approved*<br>*OMB No. 0704-0188* |
|---|---|

| | | |
|---|---|---|
| **1. REPORT DATE** *(DD-MM-YYYY)*<br>18 OCT 2007 | **2. REPORT TYPE**<br>FINAL REPORT | **3. DATES COVERED** *(From - To)*<br>01 APR 04 - 31 MAR 07 |

**4. TITLE AND SUBTITLE**

INTEGRATED RISK-SENSITIVE, SIMULATION-BASED AND GRAPHICAL METHODOLOGIES FOR ESTIMATION AND CONTROL

**5a. CONTRACT NUMBER**

**5b. GRANT NUMBER**

FA9550-04-1-0210

**5c. PROGRAM ELEMENT NUMBER**

**6. AUTHOR(S)**

Steven I. Marcus

**5d. PROJECT NUMBER**

**5e. TASK NUMBER**

**5f. WORK UNIT NUMBER**

**7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)**

Institute For Systems Research &
Electrical and Computer Engineering Dept
A.V. Williams Bldg (rm 2219),University of Maryland ,College Park, MD 20742

**8. PERFORMING ORGANIZATION REPORT NUMBER**

**9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)**

AFOSR/NL
875 NORTH RANDOLPH STREET
SUITE 325, ROOM 3112
ARLINGTON, VA 2203-1768

Dr Donald Hearn/NL

**10. SPONSOR/MONITOR'S ACRONYM(S)**

**11. SPONSOR/MONITOR'S REPORT NUMBER(S)**

**12. DISTRIBUTION/AVAILABILITY STATEMENT**

APPROVE FOR PUBLIC RELEASE, DISTRIBUTION UNLIMITED.

AFRL-SR-AR-TR-07-0508

**13. SUPPLEMENTARY NOTES**

**14. ABSTRACT**

In support of the second task, the researchers made progress incorporating simulation-based optimization and population-based methods into optimization problems. They made significant progress on new simulation-based global optimization methods, as well as on evolutionary approaches to solving Markov Decision Processes (MDPs), new sampling methods for MDPs, simulationbased methods for MDPs, new approaches to the allocation of simulation replications for optimization, and applications of these algorithms.

**15. SUBJECT TERMS**

| **16. SECURITY CLASSIFICATION OF:** | | | **17. LIMITATION OF ABSTRACT** | **18. NUMBER OF PAGES** | **19a. NAME OF RESPONSIBLE PERSON** |
|---|---|---|---|---|---|
| **a. REPORT** | **b. ABSTRACT** | **c. THIS PAGE** | | | Steven I. Marcus |
| U | U | U | | | **19b. TELEPHONE NUMBER** *(Include area code)* |

# Final Performance Report
# Grant FA95500410210

04/01/2004-03/31/2007

Steven I. Marcus, Michael C. Fu, and Alan S. Willsky

June 21, 2007

### Abstract

The researchers made significant progress in all of the proposed research areas. The first major task in the proposal involved risk-sensitive control and estimation. In support of this task, the researchers made strides toward a deeper understanding of risk-sensitive estimation and Markov chains.

In support of the second task, the researchers made progress incorporating simulation-based optimization and population-based methods into optimization problems. They made significant progress on new simulation-based global optimization methods, as well as on evolutionary approaches to solving Markov Decision Processes (MDPs), new sampling methods for MDPs, simulation-based methods for MDPs, new approaches to the allocation of simulation replications for optimization, and applications of these algorithms.

In support of the third major task that involves estimation and control algorithms for graphical models and networked systems, the researchers made progress on developing scalable algorithms for inference on graphical models. In particular, they developed a new framework for distributed, dynamic tracking and data association for multiple targets from multiple, distributed sensing nodes. This new method exploits both their communications-sensitive algorithm and their method of Nonparametric Belief Propagation.

# 1 Introduction

In this research project, we proposed to investigate integrated risk-sensitive, simulation-based, population-based and graphical methodologies for planning, estimation, and control that can be effective tools in an integrated approach to Global Awareness (Intelligence, Surveillance and Reconnaissance, or ISR) and Command and Control (C2). The questions we investigated were motivated by future Air Force requirements, which will involve a flexible and world-responsive set of missions. Issucs that arisc in this context include the fact that information for both training and operations may arrive just in time (in fact, perhaps even as forces are being deployed), requiring a much more agile, responsive, and integrated ISR-C2 system. A key idea is that, rather than separate ISR and C2 planning and execution cycles and collection managers dedicated to certain assets, there should be dynamic feedback between the ISR and C2 systems (i.e., between commanders and collection managers), and there should be dynamic allocation of sensing, collection, and processing assets.

Such systems are exceedingly complex, and we combined four approaches in the study of such problems:

- Utilizing risk-sensitive cost functions to achieve robustness and incorporate risk;
- Using simulation and other numerical methods for sequential decision making under uncertainty;
- Developing and studying efficient simulation-based and sampling methodologies for global optimization problems;
- Utilizing the structure of the system (in particular, graphical and networked models) to design scalable fast algorithms for planning, estimation, and control.

Graphical models represent a powerful framework capable of capturing spatio-temporal and hierarchical/multi-granularity relationships (cf. the Report of the Tri-Service Working Group on the Role of Probability and Statistics in Command and Command and Control [78] (Principal Authors: Prof. Alan S. Willsky, Prof. Steven I. Marcus, and Dr. Wendy Poston). The exploitation of structure, such as that inherent in graphical models, is essential for the computations involved in estimation and planning to be feasible. It has been our intention to integrate the risk-sensitive, simulation-based, population-based and computational methods discussed below with work on the control and estimation of systems described via graphical models, such as trees, dynamic Bayes networks, networked systems, and Petri nets.

Simulation and sampling can be effective tools for the analysis, design, and control of such systems (e.g., Andradóttir 1998, Jacobson and Schruben 1989, Fu 2002ab). Even those problems that can in principle be modeled using analytical techniques such as Markov chains may lead to computationally intractable models. A typical example of this is a large queueing network with general arrival processes and service time distributions, for which a simulation model can be easily and quickly built.

The need to address sequential decision making under uncertainty has led to the recent research focus on simulation-based methods for solving Markov decision processes (MDPs), which provide a useful framework for formulating these types of problems (e.g., Bertsekas and Tsitsiklis 1996, Sutton and Barto 1998). Most of the approaches have concentrated on approximating the value function, in effect reducing the dimensionality of the state space to a manageable number through a suitable parameterization (e.g., Das et al. 1999, Van Roy and Tsitsiklis 2001). The function approximation is carried out via a number of different techniques, including the use of basis functions and neural networks, where simulation is used to provide samples in order to fit curves. The key idea throughout is to avoid enumerating the entire state space. The approaches studied in this research are meant to complement these highly successful techniques.

The goal of global optimization is to find parameter values that achieve the optimum of an objective function. In general, due to the presence of multiple local optimal solutions, global optimization problems are extremely difficult to solve exactly. Solution methods for both continuous and combinatorial global optimization problems can be categorized as being either *instance-based* or *model-based* (cf. Zlochin 2004). In *instance-based* methods, the searches for new candidate solutions depend explicitly on previously generated solutions. Some well-known approaches are simulated annealing (SA) (Kirkpatrick 1983), genetic algorithms (GAs) (Srinivas 1994), tabu search (Glover 1990), and the recently proposed nested partitions (NP) method (Shi 2000).

*Model-based* search methods are a new class of solution techniques introduced in recent years. In *model-based* algorithms, new solutions are generated via an intermediate probabilistic model that is updated or induced from the previously generated solutions. In general, most of the algorithms that fall in this category share a similar framework and involving the following two phases:

1. Generate candidate solutions (random samples, trajectories) according to a specified probabilistic model (e.g., a parameterized probability distribution on the solution space).

2. Update the probabilistic model, on the basis of the data collected in the previous step, in order to bias the future search toward "better" solutions.

Some well established techniques that belong to the *model-based* methods are the cross-entropy (CE) method (Deboer 2005, Mannor 2003, Rubinstein 1997, Rubinstein 1999, Rubinstein 2001, Rubinstein2004), a class of algorithms called the estimation of distribution algorithms (EDAs) (Larranaga 1999, Muhlenbein1996, Pelikan 1999), and the so-called annealing adaptive search (AAS) (Shen 2005, Zabinsky 2003). Key questions in model-based methods are: (i) how to efficiently update the probability distributions, and (ii) how to efficiently sample from the probability distributions. While many solution techniques have been proposed, there is still a need for efficient algorithms that are based on a precise mathematical framework, are provably convergent, converge to an optimal solution quickly, are easy to implement, and handle both continuous and combinatorial, deterministic and stochastic optimization problems.

Our approach has been based on the following key points:

- Scalable fast algorithms will only be possible if one can exploit the inherent structure of the system.

- New and efficient approaches for incorporating rigorous simulation and statistical methods are required for the solution of difficult optimization and sequential decision making problems.

- Often it is much easier and more efficient to simulate complex systems than to model them analytically.

- Risk-sensitive objective functions are an effective approach for incorporating risk and achieving robustness.

## 2 Simulation-based and Sampling Methods for Global Optimization

We have considered the following optimization problem:

$$x^* \in \arg\max_{x\in\mathcal{X}} H(x), \quad x \in \mathcal{X} \subseteq \Re^n, \tag{1}$$

where the solution space $\mathcal{X}$ (which can be either continuous or discrete) is a non-empty subset of $\Re^n$, and $H(\cdot) : \mathcal{X} \to \Re$ is a deterministic function that is bounded from below, i.e., $\exists\ \mathcal{M} > -\infty$ such that $H(x) \geq \mathcal{M}\ \forall\ x \in \mathcal{X}$. We assume that problem (1) has a unique global optimal solution, i.e., there exists $x^* \in \mathcal{X}$ such that $H(x) < H(x^*)$ for all $x \neq x^*$, $x \in \mathcal{X}$, but there may be many local optimal solutions.

In Hu, Marcus, and Fu (2006b), we presented a general model-based global optimization framework called model reference adaptive search (MRAS). The motivation behind MRAS is to use a *sequence* of intermediate *reference* distributions to facilitate and guide the updating of the parameters associated with the family of parameterized distributions during the search process. The sequence of reference distributions in MRAS are selected such that they can be shown to converge to a degenerate distribution concentrated only on the set of optimal solutions. The sequence of reference models is only used *implicitly* to guide the parameter updating procedure, in contrast to the usual Estimation of Distribution Algorithms (EDAs), where the distributions must be constructed explicitly. At each iteration of MRAS, candidate solutions are generated from the distribution (among the prescribed family of distributions) that possesses the minimum Kullback-Leibler (KL) divergence with respect to the reference model corresponding to the previous iteration. These candidate solutions are in turn used to construct the next distribution that has the minimum KL-divergence with respect to the current reference model, from which future candidate solutions will be generated. For a class of parameterized probability distributions, the so-called Natural Exponential Family (NEF), the algorithm converges to an optimal solution with probability one.

To explain the main idea behind MRAS, we consider the following naive *model-based* approach for solving (1). Let $g_0(x) > 0\ \forall x \in \mathcal{X}$ be an initial probability density/mass function (p.d.f./p.m.f.) on the solution space $\mathcal{X}$. At each iteration $k \geq 1$, we compute a new p.d.f. by tilting the old p.d.f. $g_{k-1}(x)$ with the performance function $H(x)$ (for simplicity, here we assume $H(x) > 0\ \forall x \in \mathcal{X}$), i.e.,

$$g_k(x) = \frac{H(x) g_{k-1}(x)}{\int_{\mathcal{X}} H(x) g_{k-1}(dx)}, \quad \forall x \in \mathcal{X}, \tag{2}$$

By doing so, we are assigning more weight to the solutions that have better performance. One direct consequence of this is that each iteration of (2) improves the expected performance. To be precise, let $X = (X_1, \ldots, X_n)$ be a random variable taking values in $\mathcal{X}$. To reduce the notational burden, henceforth $X$ will be used to denote a random variable having the distribution under which the expectation is indicated. Thus, $E_{g_k}[H(X)] = \int_{\mathcal{X}} H(x) g_k(dx)$ and $E_{g_{k-1}}[H(X)] = \int_{\mathcal{X}} H(x) g_{k-1}(dx)$. Then we have

$$\begin{aligned} E_{g_k}[H(X)] &= \frac{E_{g_{k-1}}[(H(X))^2]}{E_{g_{k-1}}[H(X)]} \\ &\geq E_{g_{k-1}}[H(X)]. \end{aligned}$$

Furthermore, it is possible to show that the sequence of p.d.f.'s $\{g_k(\cdot),\ k = 0, 1, \ldots\}$ will converge to a p.d.f. that concentrates only on the set of optimal solutions for arbitrary $g_0(\cdot)$. So we will have $\lim_{k\to\infty} E_{g_k}[H(X)] = H(x^*)$.

However, the above approach is generally of little practical use, due to the following reasons: $(i)$ It is usually not possible to enumerate all the points in the solution space in order to perform the update (2); furthermore, if it were possible, the optimal solution could be immediately identified simply by checking which point has the best performance value. $(ii)$ The p.d.f. $g_k(x)$ constructed at each iteration may not have any structure, and therefore may be very difficult to handle.

To overcome the above difficulties, we considered in Hu, Fu, and Marcus (2006b) the Monte Carlo (sampling) version of the above approach and at the same time restrict ourselves to a family of parameterized p.d.f.'s/p.m.f.'s $\{f(\cdot,\theta)\}$, where $\theta$ is the parameter vector. In particular, at each iteration $k$ of the algorithm, we look at the projection of $g_k(\cdot)$ on the family of p.d.f.'s/p.m.f.'s $\{f(\cdot,\theta)\}$ and compute the parameter vector $\theta_k$ that minimizes the Kullback-Leibler (KL) divergence

$$\mathcal{D}(g_k, f(\cdot,\theta)) := E_{g_k}\left[\ln \frac{g_k(X)}{f(X,\theta)}\right] = \int_{x\in\mathcal{X}} \ln \frac{g_k(x)}{f(x,\theta)} g_k(x)\nu(x), \tag{3}$$

where $\nu$ is the Lebesgue/counting measure defined on $\mathcal{X}$. The benefits of the above consideration are twofold: on the one hand, $f(\cdot,\theta_k)$ often has some special structure and therefore could be much easier to handle than $g_k(\cdot)$. On the other hand, the sequence $\{f(\cdot,\theta_k)\}$ may retain some nice properties of $\{g_k(\cdot)\}$ and converge to a degenerate p.d.f. concentrated on the set of optimal solutions.

## 2.1 The MRAS$_0$ Algorithm (Exact Version)

Let $P_{\theta_k}(\cdot)$ and $E_{\theta_k}[\cdot]$ denote the probability and expectation taken with respect to the p.d.f./p.m.f. $f(\cdot,\theta_k)$, and let $I_{\{\cdot\}}$ denote the indicator function, i.e.,

$$I_{\{A\}} := \begin{cases} 1 & \text{if event } A \text{ holds,} \\ 0 & \text{otherwise.} \end{cases}$$

Thus, in this notation,

$$P_{\theta_k}(H(X) \geq \gamma) = \int_{x\in\mathcal{X}} I_{\{H(x)\geq\gamma\}} f(x,\theta_k)\nu(dx),$$

$$E_{\theta_k}[H(X)] = \int_{x\in\mathcal{X}} H(x) f(x,\theta_k)\nu(dx).$$

### Algorithm Description

The MRAS$_0$ algorithm requires specification of a parameter $\rho$, which determines the approximate proportion of samples that will be used to update the probabilistic model. At successive iterations of the algorithm, a sequence $\{\gamma_k, k = 1,2,\ldots\}$, i.e., the $(1-\rho)$-quantiles with respect to the sequence of p.d.f's $\{f(\cdot,\theta_k)\}$, are calculated at step 1 of MRAS$_0$. These quantile values are then used in step 2 to construct a sequence of non-decreasing thresholds $\{\bar{\gamma}_k, k = 1,2,\ldots\}$; and only those candidate solutions that have performances better than these thresholds will be used in parameter updating (cf. equation (4). The theoretical convergence of MRAS$_0$ is unaffected by the value of the parameter $\rho$. The purpose of $\rho$ in our approach is to concentrate the computational effort on the set of elite/promising samples, which is a standard technique employed in most of the population-based approaches, like GAs and EDAs.

During the initialization step of MRAS$_0$, a small number $\varepsilon$ and a continuous and strictly increasing function $S(\cdot) : \Re \to \Re^+$ are also specified. The function $S(\cdot)$ is used to account for the cases where the values of $H(x)$ are negative for some $x$, and the parameter $\varepsilon$ ensures that each strict increment in the sequence $\{\bar{\gamma}_k\}$ is lower bounded, i.e.,

$$\inf_{\substack{\bar{\gamma}_{k+1}\neq\bar{\gamma}_k \\ k=1,2,\ldots}} (\bar{\gamma}_{k+1} - \bar{\gamma}_k) \geq \varepsilon.$$

We require $\varepsilon$ to be strictly positive for continuous problems, and non-negative for discrete problems.

```
Algorithm MRAS_0: Model Reference Adaptive Search – exact version
• Initialization: Specify the parameter $\rho \in (0,1]$, a small number $\varepsilon \geq 0$, a
  continuous and strictly increasing function $S(\cdot) : \Re \to \Re^+$, and an initial
  p.d.f./p.m.f. $f(x, \theta_0) > 0 \ \forall x \in \mathcal{X}$. Set the iteration counter $k = 0$.
• Repeat until a specified stopping rule is satisfied:
    1. Calculate the $(1-\rho)$-quantile
           $\gamma_{k+1} := \sup_l \{l : P_{\theta_k}(H(X) \geq l) \geq \rho\}.$
    2. if $k = 0$, then set $\bar{\gamma}_{k+1} = \gamma_{k+1}$.
       elseif $k \geq 1$
           if $\gamma_{k+1} \geq \bar{\gamma}_k + \varepsilon$, then set $\bar{\gamma}_{k+1} = \gamma_{k+1}$.
           else set $\bar{\gamma}_{k+1} = \bar{\gamma}_k$.
           endif
       endif
    3. Compute the parameter vector $\theta_{k+1}$ as
           $\theta_{k+1} := \arg\max_{\theta \in \Theta} E_{\theta_k}\left[\frac{[S(H(X))]^k}{f(X,\theta_k)} I_{\{H(X) \geq \bar{\gamma}_{k+1}\}} \ln f(X,\theta)\right],$    (4)
    4. Set $k = k + 1$.
```

In continuous domains, the division by $f(x, \theta_k)$ in the performance function in step 3 is well defined if $f(x, \theta_k)$ has infinite support (e.g. normal p.d.f.), whereas in discrete/combinatorial domains, the division is still valid as long as each point $x$ in the solution space has a positive probability of being sampled. Additional regularity conditions on $f(x, \theta_k)$ ensure that step 3 of MRAS$_0$ can be used interchangeably with the following equation:

$$\theta_{k+1} = \arg\max_{\theta \in \Theta} \int_{x \in \mathcal{X}} [S(H(x))]^k I_{\{H(x) \geq \bar{\gamma}_{k+1}\}} \ln f(x,\theta) \nu(dx).$$

The following lemma shows that there is a sequence of reference models $\{g_k(\cdot),\ k = 1, 2, \ldots\}$ implicit in MRAS$_0$, and the parameter $\theta_{k+1}$ computed at step 3 indeed minimizes the KL-divergence $\mathcal{D}(g_{k+1}, f(\cdot, \theta))$.

**Lemma**. The parameter $\theta_{k+1}$ computed at the $k$th iteration of the MRAS$_0$ algorithm minimizes the KL-divergence $\mathcal{D}(g_{k+1}, f(\cdot, \theta))$, where

$$g_{k+1}(x) := \frac{S(H(x)) I_{\{H(x) \geq \bar{\gamma}_{k+1}\}} g_k(x)}{E_{g_k}\left[S(H(X)) I_{\{H(X) \geq \bar{\gamma}_{k+1}\}}\right]} \quad \forall x \in \mathcal{X}, \quad k = 1, 2, \ldots,$$

$$g_1(x) := \frac{I_{\{H(x) \geq \bar{\gamma}_1\}}}{E_{\theta_0}\left[\frac{I_{\{H(X) \geq \bar{\gamma}_1\}}}{f(X,\theta_0)}\right]}.$$

## 2.2 Global Convergence

Global convergence of the MRAS$_0$ algorithm clearly depends on the choice of parameterized distribution family. The algorithm may not be computationally tractable for some choices. In Hu, Fu, and Marcus (2006b), we have utilized a particular family of p.d.f.'s/p.m.f.'s called the natural exponential family (NEF), for which the global convergence properties can be established.

**Definition**. A parameterized family of p.d.f's $\{f(\cdot,\theta),\ \theta \in \Theta \subseteq \Re^m\}$ on $\mathcal{X}$ is said to belong to the natural exponential family (NEF) if there exist functions $h(\cdot) : \Re^n \to \Re$, $\Gamma(\cdot) : \Re^n \to \Re^m$, and $K(\cdot) : \Re^m \to \Re$ such that

$$f(x,\theta) = \exp\left\{\theta^T \Gamma(x) - K(\theta)\right\} h(x), \quad \forall \theta \in \Theta, \tag{5}$$

where $K(\theta) = \ln \int_{x\in\mathcal{X}} \exp\left\{\theta^T \Gamma(x)\right\} h(x)dx$, and the superscript "$T$" denotes the vector transposition. Many common p.d.f.'s/p.m.f.'s belong to the NEF, e.g., Gaussian, Poisson, binomial, geometric, and certain multivariate forms of them. Some regularity conditions are needed to prove convergence of the algorithm.

**Assumptions:**

**A1.** *There exists a compact set $\Pi \subseteq \mathcal{X}$ such that the level set $\{x : H(x) \geq \bar{\gamma}_1\} \subseteq \Pi$, where $\bar{\gamma}_1 = \sup_l\{l : P_{\theta_0}(H(X) \geq l) \geq \rho\}$ is defined as in the MRAS$_0$ algorithm.*

**A2.** *For any given constant $\xi < H(x^*)$, the set $\{x : H(x) \geq \xi\}$ has a strictly positive Lebesgue measure.*

**A3.** *For any given constant $\delta > 0$, $\sup_{x\in A_\delta} H(x) < H(x^*)$, where $A_\delta := \{x : \|x - x^*\| \geq \delta\}$.*

**A4.** *The maximizer of equation (4) is an interior point of $\Theta$ for all $k$.*

**A5.** $\sup_{\theta\in\Theta} \|\exp\{\theta^T\Gamma(x)\}\Gamma(x)h(x)\|$ *is integrable/summable with respect to $x$, where $\theta$, $\Gamma(\cdot)$, and $h(\cdot)$ are defined as above.*

**A6.** $\Gamma(\cdot) : \Re^m \to \Re^n$ *given above is a continuous mapping.*

**Remark** 1: Assumptions A1–A3 are regularity conditions imposed on the optimization problem to be solved, whereas assumptions A4–A6 are restrictions imposed on the parameterized family of p.d.f.'s. A1 is satisfied if the function $H(\cdot)$ has compact level sets or the solution space $\mathcal{X}$ is compact. Intuitively, assumption A2 ensures that any neighborhood of the optimal solution $x^*$ will have a positive probability of being sampled; it is satisfied if the objective function $H(\cdot)$ is continuous at $x^*$. Since $H(\cdot)$ has a unique global optimizer, A3 is satisfied by many functions encountered in practice, and is guaranteed to hold if $\mathcal{X}$ itself is compact. In actual implementation of the algorithm, step 3 of MRAS$_0$ is often posed as an unconstrained optimization problem, i.e., $\Theta = \Re^m$, in which case A4 is automatically satisfied. It is also easy to verify that A5 and A6 are satisfied by most NEFs.

**Theorem:** Let $\{\theta_k,\ k = 1, 2, \ldots\}$ be the sequence of parameters generated by MRAS$_0$. If $\varepsilon > 0$ and assumptions A1–A6 are satisfied, then

$$\lim_{k\to\infty} E_{\theta_k}[\Gamma(X)] = \Gamma(x^*). \tag{6}$$

**Remark** 2: Notice that when $\Gamma(x)$ is a one-to-one function (which is the case for many NEFs used in practice), the convergence result (6) can be equivalently written as $\Gamma^{-1}\left(\lim_{k\to\infty} E_{\theta_k}[\Gamma(X)]\right) = x^*$.

Also note that the limit in equation (6) is component-wise. For some particular p.d.f.'s/p.m.f.'s, the solution vector $x$ itself will be a component of $\Gamma(x)$ (e.g., multivariate normal distribution). Under these circumstances, we can disregard the redundant components and interpret equation (6) as $\lim_{k\to\infty} E_{\theta_k}[X] = x^*$. Another special case of particular interest is when the components of the random vector $X = (X_1, \dots, X_n)$ are independent, i.e., each has a univariate p.d.f. of the form

$$f(x_i, \vartheta_i) = \exp(x_i\vartheta_i - K(\vartheta_i))h(x_i), \ \vartheta_i \in \Re, \forall\ i = 1, \dots, n.$$

In this case, since the p.d.f. of the random vector $X$ is simply the product of the marginal p.d.f.'s, we will clearly have $\Gamma(x) = x$. Thus, equation (6) is again equivalent to $\lim_{k\to\infty} E_{\theta_k}[X] = x^*$, where $\theta_k := (\vartheta_1^k, \dots, \vartheta_n^k)$, and $\vartheta_i^k$ is the value of $\vartheta_i$ at the $k$th iteration.

## 2.3 Monte Carlo Algorithm

The MRAS$_0$ algorithm describes the idealized situation where quantile values and expectations can be evaluated exactly. In practice, we will usually resort to its stochastic (sampled, or simulation-based) counterpart, where only a finite number of samples are used and expected values are replaced with their corresponding sample averages. For example, step 3 of MRAS$_0$ will be replaced with

$$\widetilde{\theta}_{k+1} = \arg\max_{\theta\in\Theta} \frac{1}{N} \sum_{i=1}^{N} \frac{[S(H(X_i))]^k}{f(X_i, \widetilde{\theta}_k)} I_{\{H(X_i)\geq\bar{\gamma}_{k+1}\}} \ln f(X_i, \theta), \tag{7}$$

where $X_1, \dots, X_N$ are i.i.d. random samples from $f(x, \widetilde{\theta}_k)$, $\widetilde{\theta}_k$ is the estimated parameter vector computed at the previous iteration, and $\bar{\gamma}_{k+1}$ is a threshold determined by the sample $(1-\rho)$-quantile of $H(X_1), \dots, H(X_N)$.

However, the theoretical convergence can no longer be guaranteed for a simple stochastic counterpart of MRAS$_0$. In particular, the set $\{x : H(x) \geq \bar{\gamma}_{k+1}\}$ involved in equation (7) may be empty, since all the random samples generated at the current iteration may be much worse than those generated at the previous iteration. Thus, we can only expect the algorithm to converge if the expected values in the MRAS$_0$ algorithm are closely approximated. Obviously, the quality of the approximation will depend on the number of samples to be used in the simulation, but it is difficult to determine in advance the appropriate number of samples. A sample size too small will cause the algorithm to fail to converge and result in poor quality solutions, whereas a sample size too large may lead to high computational cost.

As mentioned earlier, the parameter $\rho$, to some extent, will affect the performance of the algorithm. Large values of $\rho$ mean that almost all samples generated, whether "good" or "bad", will be used to update the probabilistic model, which could slow down the convergence process. On the other hand, since a good estimate will necessarily require a reasonable amount of valid samples, the quantity $\rho N$ (i.e., the approximate amount of samples that will be used in parameter updating) cannot be too small. Thus, small values of $\rho$ will require a large number of samples to be generated at each iteration and may result in significant simulation efforts.

In order to address the above difficulties, we adopted in Hu, Fu, and Marcus (2006b) the same idea as in Homem-de-Mello and Rubinstein (2003) and proposed a modified Monte Carlo version of MRAS$_0$ in which the sample size $N$ is adaptively increasing and the parameter $\rho$ is adaptively decreasing.

### Algorithm Description and Convergence

Roughly speaking, the MRAS$_1$ algorithm is essentially a Monte Carlo version of MRAS$_0$ except that the parameter $\rho$ and the sample size $N$ may change from one iteration to another. The rate

**Algorithm MRAS$_1$: Model Reference Adaptive Search – Monte Carlo version**

```
• Initialization: Specify $\rho_0 \in (0,1]$, an initial sample size $N_0 > 1$, $\varepsilon \geq 0$,
  $\alpha > 1$, a mixing coefficient $\lambda \in (0,1]$, a continuous and strictly increasing
  function $S(\cdot) : \Re \rightarrow \Re^+$, and an initial p.d.f. $f(x,\theta_0) > 0 \; \forall x \in \mathcal{X}$. Set
  $\widetilde{\theta}_0 \leftarrow \theta_0$, $k \leftarrow 0$.
• Repeat until a specified stopping rule is satisfied:
    1. Generate $N_k$ i.i.d. samples
       $X_1^k, \ldots, X_{N_k}^k \sim \widetilde{f}(\cdot, \widetilde{\theta}_k) := (1-\lambda) f(\cdot, \widetilde{\theta}_k) + \lambda f(\cdot, \theta_0)$.
    2. Compute the sample $(1-\rho_k)$-quantile $\widetilde{\gamma}_{k+1}(\rho_k, N_k) := H_{(\lceil (1-\rho_k) N_k \rceil)}$,
       where $\lceil a \rceil$ is the smallest integer greater than $a$, and $H_{(i)}$ is the $i$th
       order statistic of the sequence $\left\{ H(X_i^k),\ i = 1, \ldots, N_k \right\}$.
    3. If $k = 0$ or $\widetilde{\gamma}_{k+1}(\rho_k, N_k) \geq \bar{\gamma}_k + \frac{\varepsilon}{2}$, then
            3a. Set $\bar{\gamma}_{k+1} \leftarrow \widetilde{\gamma}_{k+1}(\rho_k, N_k)$, $\rho_{k+1} \leftarrow \rho_k$, $N_{k+1} \leftarrow N_k$.
       else, find the largest $\bar{\rho} \in (0, \rho_k)$ such that $\widetilde{\gamma}_{k+1}(\bar{\rho}, N_k) \geq \bar{\gamma}_k + \frac{\varepsilon}{2}$.
            3b. If such a $\bar{\rho}$ exists,
                then set $\bar{\gamma}_{k+1} \leftarrow \widetilde{\gamma}_{k+1}(\bar{\rho}, N_k)$, $\rho_{k+1} \leftarrow \bar{\rho}$, $N_{k+1} \leftarrow N_k$.
            3c. else (if no such $\bar{\rho}$ exists),
                    set $\bar{\gamma}_{k+1} \leftarrow \bar{\gamma}_k$, $\rho_{k+1} \leftarrow \rho_k$, $N_{k+1} \leftarrow \lceil \alpha N_k \rceil$.
       endif
    4. Compute $\widetilde{\theta}_{k+1}$ as
       $$\widetilde{\theta}_{k+1} = \arg\max_{\theta \in \Theta} \frac{1}{N_k} \sum_{i=1}^{N_k} \frac{[S(H(X_i^k))]^k}{\widetilde{f}(X_i^k, \widetilde{\theta}_k)} I_{\{H(X_i^k) \geq \bar{\gamma}_{k+1}\}} \ln f(X_i^k, \theta). \quad (8)$$
    5. Set $k \leftarrow k+1$.
```

of increase in the sample size is controlled by an extra parameter $\alpha > 1$, specified during the initialization step. For example, if the initial sample size is $N_0$, then after $k$ increases, the sample size will be approximately $\lceil \alpha^k N_0 \rceil$.

In Hu, Fu, and Marcus (2006b), finite time $\varepsilon$-optimality, with probability 1, of this Monte-Carlo version has been proved. Numerical studies have shown that the algorithm is effective on a wide range of problems, including continuous problems with many local optima, and combinatorial problems such as asymmetric traveling salesman problems. The algorithm has also performed well on problems of topology configuration in Wave Division Multiplexed (WDM) optical networks.

## 2.4 Stochastic Model Reference Adaptive Control (SMRAS)

In Hu, Fu, and Marcus (2006c), we have extended the MRAS method to stochastic optimization problems, where the function values can only be observed in the presence of noise. Denoting $\widetilde{H}(x)$ as the random observation of the true function value $H(x)$ made at point $x$, the stochastic version of problem (1) can be formulated as

$$x^* \in \arg\max_{x \in \mathcal{X}} E[\widetilde{H}(x)], \quad x \in \mathcal{X} \subseteq \Re^n, \tag{9}$$

where $E(\cdot)$ is the expectation with respect to the probability distribution of the observation noise. Since an unbiased estimate of $E[\widetilde{H}(x)]$ is

$$\frac{1}{M}\sum_{i=1}^{M}\widetilde{H}_i(x),$$

where $\widetilde{H}_i(x),\ i=1,\dots,M$ are i.i.d. observations made at $x$, it would be natural to generalize the performance function $[S(H(x))]^k$ in MRAS to

$$S_k(\widetilde{H}(x)) := \prod_{i=1}^{k} S(\widetilde{H}_i(x)). \tag{10}$$

Clearly for the deterministic case (i.e., no observation noise) we will have the original performance function. In particular, if we take $S(\cdot)$ to be an exponential function (e.g., $S(H(x)) = e^{H(x)}$), then equation (10) can be written as

$$S_k(\widetilde{H}(x)) := \exp\Big(\sum_{i=1}^{k}\widetilde{H}_i(x)\Big).$$

Therefore, by the strong law of large numbers, it is possible to show that MRAS with the generalized performance function will converge w.p.1 to an optimal solution of (9). However, for this generalized performance function, we need to keep track of all the past observations made at all points that have been visited thus far, which could be computationally difficult to handle when the solution space is large or uncountable.

A major modification from the original MRAS method is in the way the sequence of *reference* distributions is constructed. In MRAS, *reference* distributions are *idealized* probabilistic models constructed based on the exact performance of the candidate solutions. In the stochastic case, however, the objective function cannot be evaluated deterministically, so the sample average approximations of the (idealized) *reference* distributions are used in SMRAS to guide the parameter updating. We show in Hu, Fu, and Marcus (2006c) that for the Natural Exponential Family (NEF), SMRAS converges with probability one to a global optimal solution for both stochastic continuous and discrete problems. To the best of our knowledge, SMRAS is the first *model-based* search method for solving *general* stochastic optimization problems with provable convergence. The algorithm has been shown to perform efficiently on a range of stochastic problems, including problems of buffer allocation and inventory control.

# 3 Simulation-Based and Sampling Methods for MDPs

## 3.1 Efficient Simulation Allocation via Adaptive Sampling

The basic MDP model we consider in this section is specified by the following notation:

$$\begin{aligned}
X_i &= \text{state in period } i;\\
T &= \text{time horizon, or number of periods (also known as stages)};\\
\mathcal{S} &= \text{state space};\\
\mathcal{A} &= \text{action space};\\
f_i(x,a,\omega) &= \text{transition function in period } i \text{ for action } a \text{ taken in state } x,\\
&\qquad \text{where } \omega \text{ represents the stochastic effects (e.g., a sample path)};\\
R_i(x,a,\omega) &= \text{one-period reward in period } i \text{ for action } a \text{ taken in state } x,\\
A_i \in \mathcal{A}(X_i) &= \text{action taken in period } i.
\end{aligned}$$

Thus, the MDP $\{X_i, i = 0, 1, ..., T\}$ receives reward $R_i(X_i, A_i, \omega)$ in period $i$ and then transitions according to

$$X_i = f_i(X_{i-1}, A_{i-1}, \omega).$$

The objective is to find a feedback control **policy** $\{\pi_i(x)\}_{i=0}^{T-1}$ - a mapping specifying the action taken when in state $x$ in period $i$ — that maximizes an expected reward function, which, for simplicity, we take here to be the finite horizon discounted total reward: (see Arapostathis et al. 1993):

$$E\left[\sum_{i=0}^{T-1} \alpha^i R_i(X_i, A_i, \omega)\right], \tag{11}$$

where $\alpha$ is the (one-period) discount factor; A key consideration is that simulation is required for the system dynamics (state transitions) and/or period rewards.

We define some familiar quantities:

$$\begin{aligned} Q_i(x,a) &= \text{(expected) reward-to-go ($Q$-value) in period $i$ for action $a$ taken in state $x$} \\ &\qquad \text{and optimal actions taken henceforth;} \\ J_i(x) &= \text{optimal value function in period $i$ for state $x$.} \end{aligned}$$

Then we have the usual Bellman optimality equation (e.g., Puterman 1994, Bertsekas 1995):

$$J_i(x) = \sup_a \{E[R_i(x,a) + \alpha J_{i+1}(f_{i+1}(x,a))]\}, \tag{12}$$

written here in two-part form:

$$Q_i(x,a) = E[R_i(s,a) + \alpha J_{k+1}(f_{i+1}(x,a))], \tag{13}$$

$$J_i(x) = \sup_a Q_i(x,a), \tag{14}$$

where for notational simplification, we henceforth drop explicit display of $\omega$. An optimal policy in period $i$ will be denoted by

$$\pi_i^*(x) \in \arg\sup_a Q_i(x,a), \quad i = 0, ..., T-1, \quad x \in \mathcal{S}(i). \tag{15}$$

In some applications, such as rolling-horizon control and derivatives pricing problems in finance, the goal is to estimate the optimal value function, i.e., $J_0(x_0)$ for a particular initial value $x_0$, rather than the entire optimal policy. If sampling is required to estimate the expectations involved, then the obvious way to attack the Bellman equation given by (12), or (13) and (14) is simply to replace corresponding expectation quantities with their sample means. However, given a total sampling budget, there is the question of how the budget should be allocated, both in terms of periods and in terms of actions.

To simplify the exposition in order to enhance understanding and intuition, we begin by placing some additional assumptions, primarily to reduce the notational burden. Assume that $\mathcal{A}$ is discrete and finite, so that the "sup" operation in the Bellman optimality equation becomes a "max" operation over a finite set, e.g., (12) becomes

$$J_i(x) = \max_{a \in \mathcal{A}} \{E[R_i(x,a) + \alpha J_{i+1}(f_{i+1}(x,a))]\}.$$

Again, the objective is to efficiently estimate $J_i(x)$, based on sample paths of future transitions and rewards. The estimate, along with the "best guess" for $\pi_i^*(x)$, is based on sampling over the actions $a \in \mathcal{A}$ in period $i$. In other words, our problem is how to carry out the sampling of actions from a visited state of a certain period in a sample path. We will assume that we are given a fixed $N$, the total number of samples to be distributed among the feasible actions, and $N \geq |\mathcal{A}|$, so that each action can be sampled at least once. Then the remaining question is how often should we sample each of the actions? To summarize, our problem is as follows:

- How should the sampling budget $N$ be distributed among the feasible actions (in a period)?

The simplest "solution" is what we call the equal non-adaptive scheme, in which the sampling budget is distributed equally among the feasible actions, i.e., $N/|\mathcal{A}|$ per action. So, for example, if there are ten possible actions and the sampling budget is 100, then each action would be sampled ten times to obtain sample transitions and rewards. Clearly this is generally sub-optimal, and our research is predicated on the assumption that this "equal" sampling can often lead to a tremendous waste of resources, which can be critical when the computational (sampling) budget is tight. A simple illustration of this arising in the previous example is when nine of the ten actions yield estimates of $Q(x,a)$ that are nearly deterministic, whereas the remaining action has a lot of variability (relative to all of the other actions). Then, in general, it would make much more sense to concentrate most of the sampling on the one with the high variability. Exceptions to this occur when the sample estimate of the action with high variability is worse than that of another action by an amount far exceeding that for which the variability could ever compensate; or when there is a benefit attached to sampling the best action more often.

Our approach for adaptive sampling in estimating the value function of an MDP is based on ideas from multi-armed bandit problems (cf. Gittins 1989, Berry and Fristedt 1986). The objective of these problems is to "play" (select) as often as possible the "arm", which we will call a machine henceforth, that yields the highest (expected) reward. The optimal policy must balance between playing the machine that is empirically best thus far (exploitation) — i.e., it has the highest sample mean, but not necessarily the highest expectation — and trying to find a better machine (exploration), i.e., a machine that actually has a higher expectation but might have a lower sample mean thus far due to statistical variation. Our idea is to incorporate results from this rich literature into a sampling-based process for finding an optimal action in a state for a single period of an MDP. We then extend the one-stage sampling process into multiple stages in a recursive manner, leading to a multi-stage (sampling-based) approximation algorithm for solving MDPs. Thus, by applying the theory of multi-armed bandit problems, we are able to derive a provably convergent algorithm for solving general finite-horizon MDPs.

The algorithm adaptively chooses which action to sample as the sampling process proceeds, and provides an asymptotically unbiased estimator with worst-case bias of $O\left(T\ln N/N\right)$ and worst-case time-complexity of $O\left((|\mathcal{A}|N)^T\right)$, which is independent of the size of the state space but depends on the size of the action space due to the requirement that each action be sampled *at least once* at each *reached* state.

Suppose we estimate $Q_i(x,a)$ by a sample mean $\hat{Q}_i(x,a)$ for each action $a \in \mathcal{A}$, where

$$\hat{Q}_i(x,a) = \bar{R}_i(x,a) + \alpha\frac{1}{|S^x_{a,i}|}\sum_{y\in S^x_{a,i}} \hat{J}^N_{i+1}(y), \tag{16}$$

where $S^x_{a,i}$ is the *multiset* (which means a set that may include repeated members, i.e., the same element more than once) of (independently) sampled next states from state $x$ in period $i$ taking action $a$, $|S^x_{a,i}| \geq 1$ (all actions from a state must be sampled at least once) and $\sum_{a\in A}|S^x_{a,i}| = N$ (so that the total number of sampled (next) states is $O(N^T)$, independent of the state space size), $\bar{R}_i(x,a)$ is the sample mean for the $i$th period reward, and

$$\hat{J}^N_i(x) = \sum_{a\in A}\frac{|S^x_{a,i}|}{N}\hat{Q}_i(x,a)$$

is an estimate of $J_i(x)$. This leads to the following recursion:

$$\hat{J}^N_i(x) = \sum_{a\in A}\frac{|S^x_{a,i}|}{N}\left(\bar{R}_i(x,a) + \alpha\frac{1}{|S^x_{a,i}|}\sum_{y\in S^x_{a,i}}\hat{J}^N_{i+1}(y)\right), i = 0,...,T-1,$$

with $\hat{J}_T^N(x) = 0$ for all $x \in \mathcal{S}$. Again, an adaptive scheme will specify a sequential selection of the next action to be sampled in a given state in a given period, which eventually determines $|S_{a,i}^x|$.

The main idea behind our adaptive allocation rule is based on a simple interpretation of the regret analysis of the multi-armed bandit problem, where plays of machine $m$ yield i.i.d. random rewards with unknown mean $\mu_m$, and the goal is to find/play the machine corresponding to the maximum $\mu_m$. The rewards across machines are also assumed to be independently generated. Let $T_m(n)$ be the number of times machine $m$ has been played by an algorithm during the first $n$ plays. Define the *expected regret* $\rho(n)$ of an algorithm after $n$ plays by

$$\rho(n) = \mu^* n - \sum_{m=1}^{M} \mu_m E[T_m(n)] \text{ where } \mu^* := \max_m \mu_m,$$

where $M$ is the number of possible machines. Lai and Robbins (1985) characterized an "optimal" algorithm such that the best machine, which is associated with $\mu^*$, is played exponentially more often than any other machine, at least asymptotically. That is, they showed that playing machines according to an (asymptotically) optimal algorithm leads to $\rho(n) = \Theta(\ln n)$ as $n \to \infty$ under mild assumptions on the reward distributions. Unfortunately, obtaining an optimal algorithm is often very difficult, so Agrawal (1995) derived a set of simple algorithms that achieve the asymptotic logarithmic regret behavior, using a form of *upper confidence bounds*. The temptation to play *only* the machine with the current maximum sample mean (exploitation) is tempered by the uncertainty associated with estimation, which motivates the need to play other machines occasionally (exploration). Let $\hat{\mu}_m(\bar{n})$ denote the machine $m$ sample mean, averaged over the number of plays of that machine, usually different from $\bar{n}$, which denotes the *total* (over all machines) number of plays so far. To account for the randomness in the estimation, we find a function $\sigma_m(\bar{n})$ such that $\hat{\mu}_m(\bar{n}) - \sigma_m(\bar{n}) \leq \mu_m < \hat{\mu}_m(\bar{n}) + \sigma_m(\bar{n})$ with high probability, where $\hat{\mu}_m(\bar{n}) + \sigma_m(\bar{n})$ is the upper confidence bound that guides exploration. At each play, the algorithms choose the machine with the current highest upper confidence bound.

For an intuitive description of the allocation rule, consider first only a one-stage approximation, where we assume for now that we know $J_1(x)$ for all $x \in \mathcal{S}$. Then to estimate $J_0(x)$, we need to estimate $Q_0(x, a^*)$, where $a^* \in \arg\max_a Q_0(x, a)$. The search for $a^*$ corresponds to the search for the best machine in the multi-armed bandit problem. We start by sampling each possible action once at $x$, which gives a sample one-period reward and leads to the sampled next state. We then iterate (see **Loop** in Figure 1) by sampling the action that achieves the maximum among the current estimates of $Q_0(x, a)$ plus its current upper confidence bound (see Equation (18)), where the estimate $\hat{Q}_0(x, a)$ is given by the immediate reward plus the *sample mean* of $J_1$-values at the *visited next states that have been sampled so far* (see Equation (16)). If the sampling is done appropriately, $|S_{a,0}^x|/N$ should provide a good estimate of the likelihood that action $a$ is optimal in state $x$; for $a^*$ unique, we would expect $|S_{a^*,0}^x|/N \to 1$ in the limit as $N \to \infty$. Therefore, we use a weighted (by $|S_{a,0}^x|/N$) sum of the currently estimated value of $Q_0(x, a)$ over $\mathcal{A}$ to approximate $J_0(x)$ (see Equation (19)). Ensuring that the weighted sum concentrates on $a^*$ (even if not unique) as the sampling proceeds will ensure that in the limit the estimate $\hat{J}_0^N(x)$ converges to $J_0(x)$.

We now provide a high-level description of the adaptive multi-stage sampling (AMS) algorithm given in Figure 1. The inputs to AMS are a state $x \in \mathcal{S}$, $N \geq |\mathcal{A}|$, and period $i$, and the output is $\hat{J}_i^N(x)$. AMS itself is recursively called to estimate $\hat{J}_{i+1}^N(y)$, in the **Initialization** and **Loop** subroutines of the algorithm. The initial call to AMS is done with $i = 0$ and initial state $x_0$, and every sampling is done independently of the previous samplings. To help understand how the recursive calls are made sequentially, Figure 2 graphically illustrates the sequence of calls with two actions, $\mathcal{A} = \{a, b\}$, and $T = 3$ for the **Initialization** portion. The result of this sampling scheme, as depicted in Figure 2, resemble simulated trees in the same spirit as Broadie and Glasserman (1997) use for an American-style option pricing problem and Fu and Jin (2002) use in a more general MDP setting. However, both of those works use non-adaptive sampling, in the sense that a *fixed* number of samples for each action is *pre-specified*; furthermore, all of the simulated trees are carried out in their entirety *prior to the backwards induction*. In our scheme, the sampling of actions is adaptive, and moreover the backwards induction is *integrated recursively* with the sampling.

```
Adaptive Multi-stage Sampling (AMS)
  • Input: a state $x \in \mathcal{S}$, $N \geq |A|$, and period $i$. Output: $\hat{J}_i^N(x)$, where $\hat{J}_T^N(\cdot) = 0$.
  • Initialization: Sample each feasible action $a \in \mathcal{A}$ once from state $x$ and set
        $\hat{Q}_i(x,a) = \bar{R}_i(x,a) + \alpha \hat{J}_{i+1}^N(y),$        (17)
    where $y$ is the sampled next state, and set $\bar{n} = |\mathcal{A}|$.
  • Loop: Sample each feasible action s.t.
        $a^* \in \arg\max_{a \in \mathcal{A}} \left( \hat{Q}_i(x,a) + \sqrt{\frac{2 \ln \bar{n}}{|S_{a,i}^x|}} \right),$        (18)
    where $|S_{a,i}^x|$ is the number of times action $a$ has been sampled so far,
    and $\bar{n}$ is the overall number of samples done so far for this stage.
      – Update $S_{a^*,i}^x \leftarrow S_{a^*,i}^x \cup \{y'\}$, where $y'$ is the newly visited next state from taking $a^*$.
      – Update $\hat{Q}_i(x,a^*)$ using the current $\hat{J}_{i+1}^N(y')$ according to (16).
      – $\bar{n} \leftarrow \bar{n} + 1$.
      – If $\bar{n} = N$, then go to Exit; else, continue Loop.
  • Exit: Return
        $\hat{J}_i^N(x) = \sum_{a \in \mathcal{A}} \frac{|S_{a,i}^x|}{N} \hat{Q}_i(x,a).$        (19)
```

Figure 1: Algorithm incorporating adaptive sampling.

It is not difficult to show that the time-complexity of the AMS algorithm is $O((|\mathcal{A}|N)^T)$. In contrast, the time-complexity for backward induction is $O(|\mathcal{A}||\mathcal{S}|^2T)$. Therefore, the main benefit of the proposed AMS algorithm is independence from the state space size, due to the sampling nature of the algorithm, although this comes at the expense of exponential (versus linear, for backwards induction) dependence on both the action space and the horizon length. Thus, the algorithm is most appropriate for MDPs with large state spaces but relatively small action spaces. In terms of theory, we have the following rudimentary convergence result:

**Theorem** (Chang, Hu, Fu, and Marcus 2005). For any state $x_0$,

$$\lim_{N \to \infty} E[\hat{J}_0^N(x_0)] = J_0(x_0).$$

## 3.2 Additional Methods

An alternative adaptive sampling algorithm, called the Recursive Automata Sampling Algorithm (RASA) for control of finite horizon MDPs is presented in Chang, Fu, Hu, and Marcus (2007). By extending in a recursive manner an algorithm from learning automata called the Pursuit algorithm, RASA returns estimates of both the optimal action from a given state and the corresponding optimal value. For a given initial state, we derive the following probability bounds as a function of

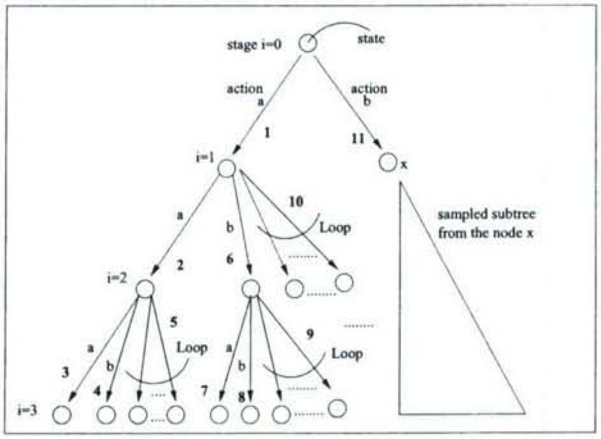

Figure 2: Graphical illustration of the sequence of the recursive calls made in **Initialization** of the AMS algorithm. Each circle corresponds to a state and each arrow with noted action signifies a sampling (and a recursive call). The bold-face number near each arrow is the sequence number for the recursive calls made. For simplicity, the entire **Loop** process is signified by one call number.

the number of samples: (i) a lower bound on the probability that RASA will sample the optimal action; and (ii) an upper bound on the probability that the deviation between the true optimal value and the RASA estimate exceeds a given error.

In recent work (Chang, Fu, and Marcus 2006), we have developed a sampling-based algorithm for solving stochastic optimization problems, based on an algorithm for solving adversarial multi-armed bandit problems. We then recursively extend the algorithm for the solution of finite horizon MDPs and analyze its performance, showing that an upper bound on the expected bias approaches zero as the sampling size per stage approaches infinity, leading to the convergence to the optimal value of the MDP.

A methodology that utilizes the approach of updating a probability distribution, but in the context of solving MDPs, has been developed in Chang, Fu, Hu, and Marcus (2006). A simulation-based algorithm, called Simulated Annealing Multiplicative Weights (SAMW), was proposed for solving large finite horizon MDPs. At each iteration of the algorithm, a probability distribution over candidate policies is updated by a simple multiplicative weight rule, and with proper annealing of a control parameter, the generated sequence of distributions converges to a distribution concentrated only on the best policies. The algorithm is asymptotically efficient, in the sense that for the goal of estimating the value of an optimal policy, a provably convergent finite-time upper bound for the sample mean is obtained.

## 4 Population-Based Evolutionary Approaches to MDPs

In this section, we discuss our research on evolutionary population-based algorithms for finding optimal (stationary) policies *infinite* horizon MDPs. These algorithms are primarily intended for problems with large (possibly uncountable) action spaces where the policy *improvement* step in Policy Iteration (PI) becomes computationally prohibitive, and value iteration is also impractical. In particular, for PI, maximizing over the entire action space may require enumeration or random search methods. The computational complexity of each iteration of our algorithms is polynomial in the size of the state space, but unlike PI and Value Iteration (VI), it is insensitive to the size of the action space, making the algorithms most suitable for problems with relatively small state spaces compared to the size of the action spaces. In the case of uncountable action spaces, our approach avoids the need for any discretization; discretization can lead to computational difficulties, either resulting in an action space that is too large or in a solution that is not accurate enough.

The approach taken by our algorithms directly searches the policy space to avoid carrying out an optimization over the entire action space at each PI step, and resembles that of a standard genetic algorithm (GA), updating a *population* of policies using appropriate analogous operations for the MDP setting. One key feature of the algorithms presented here is the determination of an elite policy that is superior to the performances of all policies in the previous population. This monotonicity property ensures that the algorithms converge with probability one to a population in which the elite policy is an optimal policy.

In Chang, Lee, Fu and Marcus (2005), we proposed a novel algorithm called Evolutionary Policy Iteration (EPI) for solving infinite horizon discounted reward MDPs. EPI inherits the spirit of the policy iteration (PI) algorithm but eliminates the need to maximize over the entire action space in the policy improvement step by directly manipulating policies via a method called "policy switching" that generates an improved policy from a set of given policies, with a computation time on the order of the size of the state space. EPI iteratively generates a population (or set) of policies such that the performance of the elite policy" for a population is monotonically improved with respect to a defined fitness function. Each iteration of the algorithms consists of two main steps: generation of an elitist policy by policy switching, and exploration of the policy space by generating additional policies via mutation and policy switching. The algorithm converges to a population that contains an optimal policy, independent of the initial population

This work is extended in Hu, Fu, Ramezani, and Marcus (2006), where a new randomized search method called Evolutionary Random Policy Search (ERPS) is introduced; ERPS considerably enhances the EPI algorithm to allow it to be more efficient for practical problems. The ERPS algorithm approaches an MDP by iteratively dividing it into a sequence of smaller, random, sub-MDP problems based on information obtained from random sampling of the entire actions space and local search, to extract a convergent sequence of policies via solving these smaller problems. It thus improves upon both the elitist policy determination and the mutation step by solving a sequence of sub-MDP problems defined on smaller policy spaces. Each sub-MDP is then solved approximately by using a variant of PI, where an elite policy is obtained.

As in EPI, each iteration of ERPS has two main steps:

1. An elitist policy is generated by solving the sub-MDP problem constructed in the previous iteration using a variant of the policy improvement technique called policy improvement with cost swapping (PICS).

2. Based on the elitist policy, a group of policies is then obtained by using a "nearest neighbor" heuristic and random sampling of the entire action space, from which a new sub-MDP is created by restricting the original MDP problem (e.g., cost structure, transition probabilities) to the current available subsets of actions. The "nearest neighbor" heuristic provides a local search mechanism that leads to rapid convergence once a policy is found in a small neighborhood of an optimal policy.

Whereas EPI treats policies as the most essential elements in the action optimization step, and each "elite" policy is directly generated from a group of policies, in ERPS policies are regarded as intermediate constructions from which sub-MDP problems are then constructed and solved. This modification substantially improves the performance while maintaining the computational complexity at essentially the same level. It is proved that the sequence of elite policies converges to an epsilon-optimal policy with probability one, and numerical studies are used to compare ERPS to other algorithms.

## 5 Risk-Sensitive Control and Estimation

In Ramezani and Marcus (2005), we have viewed the probability distribution of a Markov chain as the information state of an additive optimization problem. This optimization problem is then generalized to a product form whose information state gives rise to a generalized notion of probability distribution for Markov chains. The evolution and the asymptotic behavior of this generalized or risk-sensitive probability distribution is studied, and a conjecture is proposed regarding the asymptotic periodicity of risk-sensitive probability and is proved in the two dimensional case.

Product estimators for partially observed Markov chains are introduced in Ramezani, Marcus, and Fu (2004), and a notion of risk-sensitivity for which the risk is non-uniform and state-dependent is defined. Product probability is introduced and studied in the context of left-to-right Markov chains for uniform and state-dependent cases. It is shown that the qualitative behavior of these estimators is related to certain threshold properties. In Ramezani, Fu, and Marcus (2005), we consider the relationship between risk-sensitivity and information. Product estimators are introduced as a generalization of Maximum A Posteriori Probability (MAP) estimator for Hidden Markov Models. We study the relationship between the inclusion of higher order moments, the underlying dynamics and the availability of information. Asymptotic periodicity of these estimators and the relationship between risk and information is studied via simulation.

## 6 Optimization, Estimation, and Control in Graphical Models and Networked Systems

We made considerable progress in our work on scalable algorithms for inference in graphical models (Chen, Cetin and Willsky 2005a; Chen, Cetin, and Willsky 2005b; Chen, Wainwright, Cetin, and Willsky 2006; Ihler, Fisher, Moses, and Willsky 2005; Ihler, Fisher, and Willsky 2006; Johnson, Malioutov, and Willsky 2006). One of the applications of our methodologies that we have explored is that of multisensor, multitarget data association, a notoriously complex problem. We have now demonstrated that our new algorithms can yield remarkably efficient solutions to optimal data association problems that have heretofore been considered too complex for practical solution (hence requiring the use of heuristics to obtain tractable, but suboptimal, solutions). In addition, with an eye toward implementation in distributed sensor networks, we have developed a local, adaptive version of these data association algorithms in which, at each iteration, each node in the network can decide whether to transmit a message to each of its immediate neighbors based on whether the potential new message differs in a statistically significant manner from the previous message that was sent to that neighbor. We have shown that this locally adaptive algorithm can result in dramatic reductions in computationsand communications, if these messages were indeed sent through a sensor networkwith minimal decrease in association performance.

We have developed a new approach to inference for graphical models that involve non-Gaussian densities–problems of particular importance for various sensing modalities that provide measurements of either bearing or range. These methods, which involve the use of methods for nonparametric density estimation (for which reason we refer to them as Nonparametric Belief Propagation (NBP) algorithms), can be viewed as extensions of concepts of particle filtering to inference on graphs–this extension is highly nontrivial, especially for graphs with loops, as the iterative computations and generation of messages of belief propagation require new ideas for generating particles to replace those messages. In addition to developing the basic methodology, we have also explored applications in both computer vision and in fusion for sensor networks.

We have completed a study of the communications cost/estimation accuracy tradeoff for particle-based representations such as those used in NBP. This work provides a systematic approach to fully adaptive algorithms that directly tradeoff accuracy in the transmitted particle-based message for the total communications requirement associated with that message. Combining this with our work on relating errors between exact and transmitted messages and overall estimation accuracy, this is now the first available methodology for directly trading off overall network estimation accuracy versus communications requirements.

We have also made considerable advances in understanding inference for Gaussian graphical models, developing both very powerful, scalable, and accurate methods for covariance calculation for very large problems and also developing a new framework for analyzing and understanding distributed message-passing algorithms (based on the idea of so-called walk-sums) that provide easily computable sufficient conditions, as well as complete necessary and sufficient conditions for convergence of the well-known Belief Propagation (BP) algorithm. This perspective also suggests ways in which to achieve better performance than BP through more effective exploitation of local memory and computation in a distributed fusion system.

# 7 Additional Research Progress

We have also made significant progress in the following areas:

- Optimal allocation of simulation budget in simulation-based optimization (Chen et. al. 2004a, Fu et. al. 2004, Fu et. al. 2006);
- New results in zero-sum Markov games (Chang and Fu 2004);
- Applications in inventory control, telecommunications, preventive maintenance and production control (Zhang and Fu 2005, Chen et. al. 2004b, Ridley et. al. 2004, Yao et. al. 2006).

# 8 Research Output

## 8.1 Journal Publications

- H.S. Chang, M.C. Fu, J. Hu, and S.I. Marcus, "An Adaptive Sampling Algorithm for Solving Markov Decision Processes," Operations Research, 53, January-February 2005, 126-139.
- A.T. Ihler, J.W. Fisher, III, R.L. Moses, and A.S. Willsky, "Nonparametric Belief Propagation for Self-Localization of Sensor Networks," IEEE J. on Select Areas in Communication, special issue on Distributed Collaborative Sensor Networks, 23, April 2005, 809-819.
- V. Ramezani and S.I. Marcus, "Risk Sensitive Probability for Markov Chains," Systems and Control Letters, 54, May 2005, 493-502.
- X. Yao, X. Xie, M.C. Fu, and S.I. Marcus, "Optimal Joint Preventive Maintenance and Production Policies," Naval Research Logistics, 52, October 2005, 668-681.
- H.S. Chang, H.G. Lee, M.C. Fu, and S.I. Marcus, "Evolutionary Policy Iteration for Solving Markov Decision Processes," IEEE Transactions on Automatic Control, 50, November 2005, 1804-1808.

- S.B. Laprise, M.C. Fu, S.I. Marcus, and A.E.B. Lim, "Pricing American-Style Derivatives with European Call Options," Management Science, 52, January 2006, 95-110.

- L. Chen, M.J. Wainwright, M. Cetin, and A.S. Willsky, "Data Association Based on Optimization in Graphical Models with Application to Sensor Networks," invited paper in special issue of Mathematical and Computer Modeling on Optimization and Control for Military Applications, 43, May 2006, 1114-1135.

- C.H. Chen, D. He, and M. Fu, "Efficient Dynamic Simulation Allocation in Ordinal Optimization," IEEE Transactions on Automatic Control, 51, December 2006, 2005-2009.

- M.C. Fu, J.Q. Hu, C.H. Chen, and X. Xiong, "Simulation Allocation for Determining the Best Design in the Presence of Correlated Sampling," INFORMS Journal on Computing, 19, January 2007.

- H.S. Chang, M.C. Fu, J.Q. Hu, and S.I. Marcus, "An Asymptotically Efficient Simulation-Based Algorithm for Finite Horizon Stochastic Dynamic Programming," IEEE Transactions on Automatic Control, 52, January 2007, 89-94.

- J. Hu, M.C. Fu, V. Ramezani, and S.I. Marcus, "An Evolutionary Random Policy Search Algorithm for Solving Markov Decision Processes," to appear in INFORMS Journal on Computing.

- J. Hu, M.C. Fu, and S.I. Marcus, "A Model Reference Adaptive Search Algorithm for Global Optimization," to appear in Operations Research.

- J.K. Johnson, D.M. Malioutov, and A.S. Willsky, "Walk-Sums and Belief Propagation in Gaussian Graphical Models," to appear in J. Machine Learning Research.

- H.S. Chang, M.C. Fu, J. Hu, and S.I. Marcus, "A Survey of Some Simulation-based Methods in Markov Decision Processes," to appear in Communications in Information and Systems, 7, 2007.

- A. Rawat, H. La, M. Shayman, and S.I. Marcus, "Multicast Traffic Grooming in Unidirectional SONET/WDM Ring," to appear in IEEE Journal on Selected Areas in Communications, August 2007.

- H.S. Chang, M.C. Fu, J. Hu, and S.I. Marcus, "Recursive Learning Automata Approach to Markov Decision Processes," to appear in IEEE Transactions on Automatic Control.

### 8.2 Refereed Proceedings or Book Chapters

- H. Zhang and M.C. Fu, "Sample Path Derivatives for (s, S) Inventory Systems with Price Determination," in The Next Wave in Computing, Optimization, and Decision Technologies, Bruce Golden, S. Raghavan, Edward A. Wasil, editors, Kluwer Academic Publishers, 229-246, 2005.

- M.C. Fu, "Gradient Estimation," Chapter 19 in Handbooks in Operations Research and Management Science: Simulation, S.G. Henderson and B.L. Nelson, eds, Elsevier, 575-616, 2006.

- M.C. Fu, "Sensitivity Analysis for Simulation of Stochastic Activity Networks," in Perspectives in Operations Research: Papers in Honor of Saul Gass' 80th Birthday, F.B. Alt, M.C. Fu and B.L. Golden, editors, Springer, 351-366, 2006.

- M.C. Fu, "Variance-Gamma and Monte Carlo," in Advances in Mathematical Finance, M.C. Fu, R.A. Jarrow, J.-Y. Yen, and R. J. Elliott, editors, Birkhauser, 2007.

- C.-II. Chen, D. IIe, and M.C. Fu, "A Case Study for Optimal Dynamic Simulation Allocation in Optimal Optimization," Proc. American Control Conference, 2004, 5754-5759.

- V. Ramezani, S.I. Marcus, and M.C. Fu, "Structured Risk-Sensitivity for Partially Observed Markov Chains," Proc. 43rd IEEE Conference on Decision and Control, December 2004, 3473-3478.

- H.S. Chang and M.C. Fu, "Localization for a Class of Two-Team Zero-Sum Markov Games," Proc. 43rd IEEE Conference on Decision and Control, December 2004, 4844-4849.

- M. Chen, J.Q. Hu, and M.C. Fu, "Fluid Approximation and Perturbation Analysis of a Dynamic Priority Call Center," Proc. 43rd IEEE Conference on Decision and Control, December 2004, 2304-2309.

- V. Ramezani, M.C. Fu, and S.I. Marcus, "Risk and Information in the Estimation of Hidden Markov Models," Proc. 2004 Winter Simulation Conference, December 2004, 1596-1601.

- M.C. Fu, J.Q. Hu, C.H. Chen, and X. Xiong, "Optimal Computing Budget Allocation Under Correlated Sampling," Proc. 2004 Winter Simulation Conference, December 2004, 595-603.

- P. Fard, R. J. La, K. Lee, S. I. Marcus, and M. Shayman, "Reconfiguration of MPLS/WDM Networks Using Simulation-Based Markov Decision Processes," Proc. 39th Annual Conference on Information Sciences and Systems, Baltimore, MD, February 2005.

- L. Chen, M. Cetin, and A.S. Willsky, "Graphical Model-Based Algorithms for Data Association in Distributed Sensing," Adaptive Sensor Array Processing Workshop, MIT Lincoln Laboratory, June 7-8, 2005.

- M.C. Fu, J.Hu, and S.I. Marcus, "Population-Based Evolutionary Approaches for Solving Markov Decision Processes," Proc. 2005 IFORS Conference, July 11-15, 2005, Honolulu, Hawaii.

- L. Chen, M. Cetin, and A.S. Willsky, "Distributed Data Association for Multi-Target Tracking in Sensor Networks," Intl. Conf. On Information Fusion, July 2005; Best Student Paper Award.

- C. Panayiotou, W.C. Howell, and M.C. Fu, "Online Traffic Light Control Through Gradient Estimation Using Stochastic Fluid Models," Proc. IFAC Triennial World Congress, 2005.

- M. C. Fu, "Sensitivity Analysis for Stochastic Activity Networks," Proc. International Conf. on Automatic Control and Systems Engineering, 2005.

- Y. He, M. C. Fu, and S. I. Marcus, "A Two-Timescale Simulation-Based Gradient Algorithm for Weighted Cost Markov Decision Processes," Proc. 44th IEEE Conference on Decision and Control, December 2005, 8022-8027.

- H. S. Chang, M. C. Fu, and S. I. Marcus, "Recursive Learning Automata for Control of Partially Observable Markov Decision Processes," Proc. 44th IEEE Conference on Decision and Control, December 2005, 6091-6096.

- R. L. Bennett, M. C. Fu, R. Jarrow, D.A. Nuxoll, and H. Zhang, "A Loss Default Simulation Model of the Federal Bank Deposit Insurance Funds," Proc. 2005 Winter Simulation Conference, December 2005, 1835-1843.

- M. C. Fu, F.W. Glover, and J. April, "Simulation Optimization: A Review, New Developments, and Applications," Proc. 2005 Winter Simulation Conference, December 2005, 83-95.

- J. Hu, M. C. Fu, and S. I. Marcus, "Stochastic Optimization using Model Reference Adaptive Search," Proc. 2005 Winter Simulation Conference, December 2005, 811-818.

- J.K. Johnson, D.M. Malioutov, and A.S. Willsky, "Low-Rank Variance Estimation in Large-Scale GMRF Models, ICASSP 2006, Toulouse, France; winner Outstanding Student Paper Award.

- M.C. Fu, J. Hu, and S.I. Marcus, "Model-Based Randomized Methods for Global Optimization," Proc. 17th International Symposium on Mathematical Theory of Networks and Systems, Kyoto, Japan, July 24-28, 2006, 355-363.

- H. Zhang and M. C. Fu, "Applying Model Reference Adaptive Search to American-Style Option Pricing," Proc. 2006 Winter Simulation Conference, December 2006, 711-718.

- H. S. Chang, M. C. Fu, and S. I. Marcus, "Adversarial Multi-Armed Bandit Approach to Stochastic Optimization," Proc. 45th IEEE Conference on Decision and Control, December 2006, 5681-5686.

- Y. Xin, M. Shayman, R.J. La, and S.I. Marcus, "Reconfiguration of Survivable MPSL/WDM Networks," Proc. IEEE GLOBECOM , San Francisco, CA, Nov. 27-Dec. 1, 2006.

- M. C. Fu and W.C. Howell, "Traffic Light Signal Optimization via Simulation," Proc. International Modeling and Simulation Multiconference: AI, Simulation and Planning in High Autonomy Systems (AIS) and Conceptual Modeling and Simulation (CMS), 2007.

### 8.3 Authored Books or Monographs

- H.S. Chang, M.C. Fu, J. Hu, and S.I. Marcus, Simulation-based Algorithms for Markov Decision Processes, Springer-Verlag, 2007 (research monograph).

### 8.4 Edited Volumes

- F.B. Alt, M.C. Fu and B.L. Golden, editors, Perspectives in Operations Research: Papers in Honor of Saul Gass' 80th Birthday, Springer-Verlag, 2006.

- M.C. Fu, R.A. Jarrow, J.-Y. Yen, and R. J. Elliott, editors, Advances in Mathematical Finance, Birkhauser, 2007.

### 8.5 Awards

- Alan Willsky: 2004 IEEE Donald G. Fink Prize Paper Award
- Alan Willsky: Univ. de Rennes, Doctorate Honoris Causa, 2005
- Best Student Paper Award: L. Chen, M. Cetin, and A.S. Willsky, Distributed Data Association for Multi-Target Tracking in Sensor Networks, Intl. Conf. On Information Fusion, July 2005.
- Outstanding Student Paper Award: J.K. Johnson, D.M. Malioutov, and A.S. Willsky, Low-Rank Variance Estimation in Large-Scale GMRF Models, ICASSP 2006, Toulouse, France

### 8.6 Ph.D. Students

- Martin Wainwright, Ph.D, 2005, MIT, supervised by A. Willsky, currently Assistant Professor in EECS and Statistics, Univ. of California, Berkeley
- Alex Ihler, Ph.D, 2005, MIT, supervised by A. Willsky, currently Assistant Professor, Toyota Technical Institute, Chicago
- Jiaqiao Hu, Ph.D, 2006, Univ. of Maryland, supervised by S. Marcus and M. Fu; currently an Assistant Professor in Applied Mathematics and Statistics at SUNY Stony Brook
- Enlu Zhou, Ph.D expected 2008, Univ. of Maryland, supervised by S. Marcus and M. Fu
- Yongqiang Wang, Ph.D expected 2010, Univ. of Maryland, supervised by S. Marcus and M. Fu
- Lei Chen, Ph.D expected May 2007, MIT, supervised by A. Willsky
- Dmitry Malioutov, Ph.D expected 2008, MIT, supervised by A. Willsky

### 8.7 Postdocs

- Vahid Ramezani, 2004-2005, Univ. of Maryland, supervised by S. Marcus and M. Fu; currently at IAI
- Ying He, 2004-2005, Univ. of Maryland, supervised by S. Marcus and M. Fu; currently at NIH

## References

[1] R. Agrawal, "Sample Mean-based Index Policies with $O(\log n)$ Regret for the Multi-armed Bandit Problem," *Advances in Applied Probability*, **27**, 1054-1078, 1995.

[2] R. Agrawal, D. Teneketzis, and V. Anantharam, "Asymptotically Efficient Adaptive Allocation Schemes for Controlled Markov Chains: Finite Parameter Space," *IEEE Trans. on Automatic Control*, **34**, 1249-1259, 1989.

[3] P. Auer, N. Cesa-Bianchi, and P. Fisher, "Finite-time Analysis of the Multiarmed Bandit Problem," *Machine Learning*, **47**, 235-256, 2002.

[4] S. Andradóttir, "Simulation Optimization," Chapter 9 in *Handbook of Simulation: Principles, Methodology, Advances, Applications, and Practice*, Wiley, 1998.

[5] A. Arapostathis, V.S. Borkar, E. Fernández-Gaucherand, M.K. Ghosh and S.I. Marcus, "Discrete-Time Controlled Markov Processes with Average Cost Criterion: A Survey, *SIAM Journal on Control and Optimization*, **31**, 282-344, 1993.

[6] J. Bean, W. Hopp, and I. Duenyas, "A Stopping Rule for Forecast Horizon in Nonhomogeneous Markov Decision Processes, *Operations Research*, **40**, 1188-1199, 1992.

[7] R.E. Bechhofer, T.J. Santner, and D.M. Goldsman, *Design and Analysis of Experiments for Statistical Selection, Screening, and Multiple Comparisons*, Wiley, 1995.

[8] D.A. Berry and B. Fristedt, *Bandit Problems: Sequential Allocation of Experiments*, Routledge, 1986.

[9] D.P. Bertsekas, *Dynamic Programming and Optimal Control, Vol. 1 & 2*, Athena Scientific, 1995.

[10] D.P. Bertsekas and J. N. Tsitsiklis, *Neuro-Dynamic Programming*, Athena Scientific, 1996.

[11] M. Broadie and P. Glasserman, "Pricing American-Style Securities Using Simulation," *Journal of Economic Dynamics and Control*, **21**, 1323-1352, 1997.

[12] H.S. Chang and M.C. Fu, "Localization for a Class of Two-Team Zero-Sum Markov Games,?*Proceedings of the 43rd IEEE Conference on Decision and Control*, December 2004, 4844-4849.

[13] H.S. Chang, M.C. Fu, J.Q. Hu, and S.I. Marcus, "An Asymptotically Efficient Simulation-Based Algorithm for Finite Horizon Stochastic Dynamic Programming," to appear in *IEEE Transactions on Automatic Control*, 2006.

[14] H.S. Chang, M.C. Fu, J.Q. Hu, and S.I. Marcus, "Recursive Learning Automata Approach to Markov Decision Processes," to appear in *IEEE Transactions on Automatic Control*, 2007.

[15] H.S. Chang, M.C. Fu, and S.I. Marcus, "An Asymptotically Efficient Simulation-based Algorithm for Finite Horizon Stochastic Dynamic Programming," *Proceedings of the 42nd IEEE Conference on Decision and Control*, 3818-3823 (2003).

[16] H.S. Chang, M.C. Fu, and S.I. Marcus, "Adversarial Multi-Armed Bandit Approach to Stochastic Optimization," *Proceedings of the 45nd IEEE Conference on Decision and Control*, (2006).

[17] H.S. Chang, J.Q. Hu, M.C. Fu, and S.I. Marcus, "An Adaptive Sampling Algorithm for Solving Markov Decision Processes," *Operations Research*, **53**, 126-139, January-February 2005.

[18] H.S. Chang, H.-G. Lee, M.C. Fu, and S.I. Marcus, "Evolutionary Policy Iteration for Solving Markov Decision Processes," *IEEE Transactions on Automatic Control*, **50**, 1804-1808, November 2005.

[19] C.-H. Chen, D. He, and M.C. Fu, "A Case Study for Optimal Dynamic Simulation Allocation in Optimal Optimization," *Proceedings of the American Conference on Control*, 5754-5759, 2004a.

[20] L. Chen, M. Cetin, and A.S. Willsky, "Distributed Data Association for Multi-Target Tracking in Sensor Networks," *Proceedings of the Intl. Conf. On Information Fusion*; Best Student Paper Award, 2005a.

[21] L. Chen, M. Cetin, and A.S. Willsky, "Graphical Model-Based Algorithms for Data Association in Distributed Sensing," Adaptive Sensor Array Processing Workshop, MIT Lincoln Laboratory, 2005b.

[22] L. Chen, M.J. Wainwright, M. Cetin, and A.S. Willsky, "Data Association Based on Optimization in Graphical Models with Application to Sensor Networks," invited paper in the special issue of *Mathematical and Computer Modeling on Optimization and Control for Military Applications* (Juan Vasquez, ed.), **43**, 1114-1135, May 2006.

[23] M. Chen, J.Q. Hu, and M.C Fu, "Fluid Approximation and Perturbation Analysis of a Dynamic Priority Call Center," *Proceedings of the 43rd IEEE Conference on Decision and Control*, 2304-2309, 2004b.

[24] W.L. Cooper, S.G. Henderson, and M.E. Lewis, "Convergence of Simulation-Based Policy Iteration," *Probability in the Engineering and Informational Sciences*, 2002.

[25] L. Dai, "Convergence Properties of Ordinal Comparison in the Simulation of Discrete Event Dynamic Systems," *Journal of Optimization Theory and Applications,* **91**, 363-388, 1996.

[26] L. Dai and C. Chen, "Rate of Convergence for Ordinal Comparison of Dependent Simulations in Discrete Event Dynamic Systems," *Journal of Optimization Theory and Applications,* **94**, 29-54, 1997.

[27] T.K. Das, A. Gosavi, S. Mahadevan, and N. Marchalleck, "Solving Semi-Markov Decision Problems Using Average Reward Reinforcement Learning," *Management Science*, **45**, 560-574, 1999.

[28] P.T. De Boer, D.P. Kroese, S. Mannor, and R.Y. Rubinstein, "A Tutorial on the Cross-entropy Method," *Annals of Operation Research,* **134**, 19-67, (2005).

[29] M.C. Fu, "Optimization for Simulation: Theory vs. Practice" (Feature Article), *INFORMS Journal on Computing*, Vol.14, No.3, 192-215, 2002a.

[30] M.C. Fu, "Simulation Optimization: Evolution or Revolution?" *INFORMS Journal on Computing*, Vol.14, No.3, 226-227, 2002b.

[31] M.C. Fu, J.Q. Hu, C.H. Chen, and X. Xiong, "Optimal Computing Budget Allocation Under Correlated Sampling," *Proceedings of the 2004 Winter Simulation Conference*, December 2004, 595-603. .

[32] M.C. Fu, J.Q. Hu, C.H. Chen, and X. Xiong, "Simulation Allocation for Determining the Best Design in the Presence of Correlated Sampling," *INFORMS Journal on Computing*, to appear, 2006.

[33] M.C. Fu and X. Jin, "Convergence of Simulation-Based Policies for Stochastic Dynamic Programming," Technical Report, 2004.

[34] J.C. Gittins, *Multi-Armed Bandit Allocation Indices*, Wiley, 1989.

[35] R. Givan, S. Leach, and T. Dean, "Bounded Markov decision Processes," *Artificial Intelligence*, **122**, 71-109, 2000.

[36] F. Glover, "Tabu Search: A Tutorial," *Interfaces*, **20**, 74–94, (1990).

[37] D. Goldsman and B.L. Nelson, "Comparing Systems via Simulation," Chapter 9 in *Handbook of Simulation: Principles, Methodology, Advances, Applications, and Practice*, Wiley, 273-306, 1998.

[38] Y. Hochberg and A.C. Tamhane, *Multiple Comparison Procedures*, Wiley, 1987.

[39] T. Homem-de-Mello, A. Shapiro, and M.L. Spearman, "Finding Optimal Material Release Times using Simulation-Based Optimization," *Management Science*, **45**, 86-102, 1999.

[40] J. Hu, M.C. Fu, and S.I. Marcus, "Model Reference Adaptive Search: A New Approach to Global Optimization," *late-breaking paper, Genetic and Evolutionary Computation Conference (GECCO)*, Washington D. C. (2005a).

[41] J. Hu, M.C. Fu, and S.I. Marcus, "Simulation Optimization using Model Reference Adaptive Search," *Proceedings of the 2005 Winter Simulation Conference,* 811–818, (2005b).

[42] J. Hu, M.C. Fu, V. Ramezani, and S.I. Marcus, "An Evolutionary Random Policy Search Algorithm for Solving Markov Decision Processes," *INFORMS Journal on Computing,* forthcoming, (2006a).

[43] J. Hu, M.C. Fu, and S.I. Marcus, "A Model Reference Adaptive Search Algorithm for Global Optimization," *Operations Research,* forthcoming (2006b).

[44] J. Hu, M.C. Fu, and S.I. Marcus, "A Model Reference Adaptive Search Method for Stochastic Global Optimization," submitted for publication (2006c).

[45] A.T. Ihler, J.W. Fisher, III, R.L. Moses, and A.S. Willsky, "Nonparametric Belief Propagation for Self-Localization of Sensor Networks," *IEEE J. on Select Areas in Communication*, special issue on Distributed Collaborative Sensor Networks, 808-819, **23**, No. 4, 2005.

[46] A.T. Ihler, J.W. Fisher, III, and A.S. Willsky, "Using Sample-Based Representations Under Communications Constraints," *ACM Transactions on Sensor Networks*, in review, 2006.

[47] S.H. Jacobson and L.W. Schruben, "A Review of Techniques for Simulation Optimization," *Operations Research Letters*, **8**, 1-9, 1989.

[48] L.P. Kaelbling, M.L. Littman, and A.W. Moore, "Reinforcement Learning: A Survey," *Journal of Artificial Intelligence*, **4**, 237-285, 1996.

[49] S.-H. Kim and B.L. Nelson, "A Fully Sequential Procedure for Indifference-Zone Selection in Simulation," *ACM Transactions on Modeling and Computer Simulation*, **11**, 251-273, 2001.

[50] S. Kirkpatrick, C.D. Gelatt, and M.P. Vecchi, "Optimization by Simulated Annealing," *Science,* **220**, 671-680 (1983).

[51] T. Lai H. and Robbins, "Asymptotically Efficient Adaptive Allocation Rules," *Advances in Applied Mathematics*, **6**, 4-22, 1985.

[52] P. Larrañaga, R. Etxeberria, J.A. Lozano, B . Sierra, I. Iñza, and J.M. Peña, "A Review of the Cooperation between Evolutionary Computation and Probabilistic Graphical Models," *Proceedings of the Second Symposium on Artificial Intelligence. Adaptive Systems. CIMAF* 99. Special Session on Distributions and Evolutionary Computation, 314–324 (1999).

[53] J.K. Johnson, D.M. Malioutov, and A.S. Willsky, "Walk-Sums and Belief Propagation in Gaussian Graphical Models, accepted for publication in *J. Machine Learning Research.*

[54] S. Mahadevan, "Average Reward Reinforcement Learning: Foundations, Algorithms, and Empirical," *Machine Learning*, **22**, 159-195, 1996.

[55] S. Mannor, R. Rubinstein, and Y. Gat, "The Cross-entropy Method for Fast Policy Search," *International Conference on Machine Learning,* 512–519 (2003).

[56] H. Mühlenbein, and G. Paaß, "From Recombination of Genes to the Estimation of Distributions: *I.* Binary Parameters," *In Hans-Michael Voigt, Werner Ebeling, Ingo Rechenberg, and Hans-Paul Schwefel, editors, Parallel Problem Solving from Nature - PPSN IV,* 178–187, Berlin, Springer Verlag, (1996).

[57] M. Pelikan, D.E. Goldberg, and F.G. Lobo, "A Survey of Optimization by Building and using Probabilistic Models," Urbana, IL: University of Illinois Genetic Algorithms Laboratory (IlliGAL report No. 99018), (1999).

[58] G.C. Pflug, *Optimization of Stochastic Models*, Kluwer Academic, 1996.

[59] M.L. Puterman, *Markov Decision Processes*, John Wiley & Sons, New York, 1994.

[60] V. Ramezani and S.I. Marcus, "Risk Sensitive Probability for Markov Chains,?*Systems and Control Letters*, **54**, May 2005, 493-502.

[61] V. Ramezani, S.I. Marcus, and M.C. Fu, "Structured Risk-Sensitivity for Partially Observed Markov Chains," *Proc. 43rd IEEE Conference on Decision and Control*, December 2004a.

[62] V. Ramezani, M.C. Fu, and S.I. Marcus, "Risk and Information in the Estimation of Hidden Markov Models," *Proc. 2004 Winter Simulation Conference*, December 2004b.

[63] S. Rathinam, R. Sengupta, and S. Sarbha, "A Resource Allocation Algorithm for Multi-Vehicle Systems with Nonholonomic Constraints," preprint, 2005.

[64] A.D. Ridley, W. Massey, and M.C. Fu, "Fluid Approximation of a Priority Call Center with Time-Varying Arrivals," *The Telecommunications Review*, 15, 69-77, 2004.

[65] R.Y. Rubinstein, "Optimization of Computer Simulation Models with Rare Events," *European Journal of Operations Research,* **99**, 89–112, (1997).

[66] R.Y. Rubinstein, "The Cross-entropy Method for Combinatorial and Continuous Optimization," *Methodology and Computing in Applied Probability,* **2**, 127–190 (1999).

[67] R.Y. Rubinstein, "Combinatorial Optimization, Ants and Rare Events," In S. Uryasev and P. M. Pardalos, editors, Stochastic Optimization: Algorithms and Applications, 304–358 Kluwer (2001).

[68] R.Y. Rubinstein and D.P. Kroese, *The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation, and Machine Learning,* Springer, New York, 2004.

[69] R.Y. Rubinstein and A. Shapiro, *Discrete Event Systems: Sensitivity Analysis and Stochastic Optimization by the Score Function Method,* John Wiley & Sons, 1993.

[70] K. Savla, F. Bullo, and E. Frazzoli, "On Traveling Salesperson Problcms fro Dubins' Vchicle: Stochastic and Dynamic Environments," *Proceedings of the 44th IEEE Conference on Decision and Control,* Seville, Spain, December 2005, 4530-4535.

[71] Y. Shen, *Annealing Adaptive Search with Hit-and-Run Sampling Methods for Global Optimization,* Ph.D. Thesis, Department of Industrial Engineering, University of Washington, Seattle, 2005.

[72] L. Shi and S. Olafsson, "Nested Partitioned Method for Global Optimization," *Operations Research,* **48**, 390-407, 2000.

[73] A. N. Shiryayev, *Probability,* New York: Springer-Verlag, 1984.

[74] A. Smith, A. Doucet, N. de Freitas, and N. Gordon, *Sequential Monte Carlo Methods in Practice,* New York: Springer-Verlag, 2001.

[75] M. Srinivas, and L.M. Patnaik, "Genetic Algorithms: A Survey," *IEEE Comput.,* **27**(6) 17–26 (1994).

[76] R.S. Sutton and A.G. Barto, *Reinforcement Learning: An Introduction,* MIT Press, Cambridge, 1998.

[77] B. Van Roy and J. N. Tsitsiklis, "Regression Methods for Pricing Complex American-style Options," *IEEE Transactions on Neural Networks,* **14**, 694-703, 2001.

[78] A.S. Willsky, S.I. Marcus, and W. Poston, "Report of the Tri-Service Working Group on the Role of Probability and Statistics in Command and Control," Office of Naval Research, 1997.

[79] W.N. Yang and B.L. Nelson, "Using Common Random Numbers and Control Variates in Multiple-Comparison Procedures," *Operations Research,* **39**, 583-591, 1991.

[80] X. Yao, X. Xie, M.C. Fu, and S. I. Marcus, "Optimal Joint Preventive Maintenance and Production Policies", to appear in *Naval Research Logistics.*

[81] Z.B. Zabinsky, *Stochastic Adaptive Search for Global Optimization,* Kluwer Academic Publisher, Norwell, MA, 2003.

[82] H. Zhang and M.C. Fu, "Sample Path Derivatives for (s, S) Inventory Systems with Price Determination,in *The Next Wave in Computing, Optimization, and Decision Technologies,* Bruce Golden, S. Raghavan, Edward A. Wasil, editors, Kluwer Academic Publishers, 229-246, 2005.

[83] M. Zlochin, M. Birattari, N. Meuleau, and M. Dorigo, "Model-based Search for Combinatorial Optimization: A Critical Survey," *Annals of Operations Research,* **131**, 373–395, (2004).